Manuscript version
DOI original paper: 10.1109/ITSC55140.2022.9922167

Complementary and situation sensitive object detection, review performance, and performance dependencies of common approaches

Waldemar Boschmann[1] and Dirk Söffker[1]

***Abstract*— Object detection can be performed on different modalities like camera or lidar systems. Image-based approaches are highly sensitive to variations of illumination. On the other hand, lidar-based approaches fail on detecting small objects or objects at higher distances. The research in this field leads to a variety of different approaches with different advantages and drawbacks. Fusion-based approaches aim to utilize the advantages of the available modalities and specialized approaches. Usually, the performance of object detection approaches is measured over a set of situations. The performance measures are typically provided as average values over a test dataset. Local variations are not considered for comparison and not utilized. For a better understanding, this contribution analyzes performance variations based on different situation parameters to understand the complementary potential and improve the application of redundant and situated object detection systems.**

I. INTRODUCTION

Nowadays learning-based methods are gaining significant attention and are utilized in safety-critical tasks, affecting decision making due to the obtained information. While most of these approaches work well in certain situations, overreliance can still cause a significant amount of damage. Due to the approaching usage of learning-based methods in real applications, there is an increasing number of accidents based on wrong decisions during autonomous or assisted operation. A popular example is an autonomous vehicle confusing a truck with a bright sky resulting in a collision [1]. Understanding abilities and uncertainties could lead to better decision-making and therefore to a reduction of risk during the operation. However, in practice, the quantification of reliability of a particular prediction is given by a detection score, estimated by the trained model. While a higher score indicates higher confidence it does not reflect the actual uncertainty of the prediction. It remains difficult to decide whether to accept or reject a prediction due to the lack of situational knowledge and a reliable indicator of quality. Due to the performance potential and the resulting research interest, a high variation of promising approaches is available. It can be assumed that a combination of diverse approaches can compensate for each other's drawbacks and lead to better and more robust predictions, improving the reliability of predictions and final decisions. This contributin is structured as follows. In Section II a brief review on object detection, object detection fusion, and situational dependencies for the detection task is provided. In Section III, used approaches and dataset are introduced. Experimental results are disscussed in Section IV.

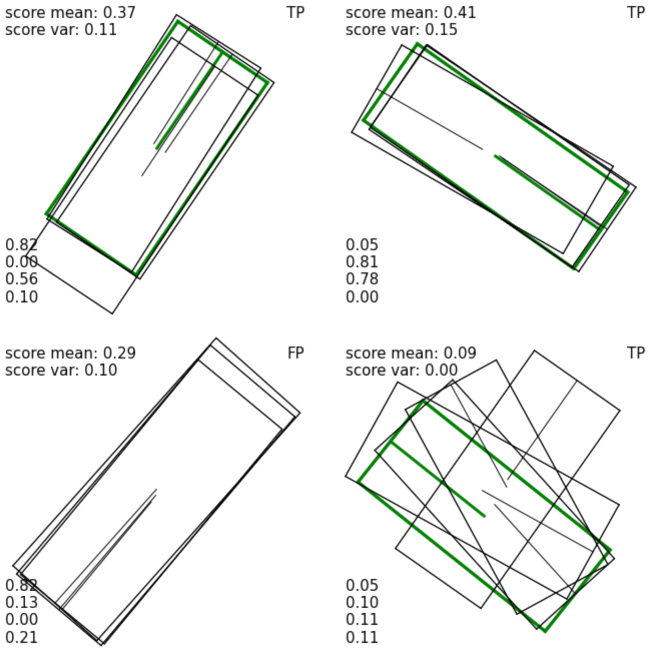

Fig. 1. Example of matched predictions of multiple detection systems. Different instances shown. Annotations for each instance as follows.
Top left: detection score mean and variance;
Top right: true or false positiv, ground truth in green;
Bottom left: detection score of matched predictions, sequence corresponds to prediction case[1] in Table III

II. RELATED WORK

A. 2D image- and 3D lidar-based detection

Image-based object detection has received high research interest over the last years and shows significant representation ability, especially for the classification task. It is well known that image-based approaches are sensitive to light conditions and textural information being provided. While achieving reasonable results under good conditions, the performance reduces significantly for varying brightness, noise, or a poor texture like repeating patterns or uniform surfaces. Furthermore estimating 3D positions based on images remains a challenging task. State-of-the-art approaches can be divided into two main categories, one-stage, and two-stage. One stage systems [2] [3] learn the detection task

We acknowledge support by the European Regional Development Fund (ERDF), grant-no. EFRE-0801714

[1]Waldemar Boschmann[1] and Dirk Söffker are with the Chair of Dynamics and Control, University of Duisburg-Essen, Duisburg, Germany, waldemar.boschmann@uni-due.de and soeffker@uni-due.de

directly from the full image and can be trained in an end-to-end fashion. This leads to a simple architecture and low inference time but shows lower performance compared to two-stage approaches. Two-stage approaches rely on region proposals that can either be precomputed [4] or provided by a region proposal network [5]. The region proposal network can either operate on a feature map, obtained from the image or on additional sensor data as demonstrated in [6]. Lidar-based object detection shows high performance for 3D object detection due to the obtained depth information provided by the sensor system. While lidar data provide precise but sparse depth, textural information is not available. The degree of available information is reduced depending on distance, object size, or shape. Therefore the detectability of an object is limited due to the obtained information. In [7] semantic segmentation is applied on raw point clouds. Based on the segmentation object detection is performed using connected components. More recent approaches transform the point cloud into voxel or pillar representations leading to a dense structure. Features are extracted voxel- or pillar wise and forwarded to a region proposal network and final detection stage [8][9][10][11].

B. Object detection and decision fusion

Fusion aims to combine existing advantages by utilizing complementary aspects of available information. Fusion can be grouped either as fusion of raw measurements (early fusion), fusion of feature maps (middle fusion), or fusion of predicted candidates (late fusion). Fusion of predicted candidates can be denoted as decision fusion. Early fusion can be performed by projections or augmentation of the available sensor data. For example, a lidar point cloud is projected into a front view depth map and used as an overlay for camera images [12]. The other way round camera images can be used to augment a point cloud by adding pixel color as additional information [13]. Middle fusion can be performed in various ways, examples are following. Nabati et al. [6] proposed the generation of region proposals based on radar points and evaluate the generated proposals with an image-based detection system like Fast R-CNN [5]. In [14] features are generated separately for images and point clouds. Proposals are generated in 3D through the lidar network. Features from lidar and image domain are presented to a fusion network performing scoring and bounding box regression using proposals and ROI features. Proposals from radar or image domain are suggested for practical application. Late fusion aims for the fusion on high level features or preliminary predictions. Qi et al. [15] used a pre-trained model to predict 2D region proposals on image data. Each proposal transformed into a frustum, limiting the search space for a lidar-based detection pipeline. While reporting improvement, this method requires detections in both domains, camera, and lidar. Pang et al. [16] fused 2D and 3D predictions by a lightweight fusion network. Fusion was performed based on the predicted score, intersection in the 2D image plane, and distance in 3D. The fusion result was a new score map. It can be assumed that the used model can learn distance dependencies. Further influences are not represented.

TABLE I
CROSS VALIDATION SCHEME - NUMBER OF SCENES ASSIGNED TO THE INDIVIDUAL FOLD

	Train			
	Rain	Night	Rain and Night	Default
LOOCV 1	125	56	8	491
LOOCV 2	82	59	14	525
LOOCV 3	149	78	11	442
LOOCV 4	108	56	15	501
LOOCV 5	132	83	16	449
	Val			
	Rain	Night	Rain and Night	Default
LOOCV 1	24	27	8	111
LOOCV 2	67	24	2	77
LOOCV 3	0	5	5	160
LOOCV 4	41	27	1	101
LOOCV 5	17	0	0	153
	17,5 %	9,8 %	1,9 %	70,8 %

C. Situational variation in object detection

Real-world applications are facing a diverse set of situations. Situations are defined by environmental influences introducing uncertainties. These environmental influences are represented in the sensor data as well as the obtained predictions of single or multiple detection systems. Influences like weather, illumination of the scene, or similar, can affect the whole detection range. Quantification can be done independent of predictions. Influences represented in the prediction like distance, confidence or associated ROI features affect a particular prediction. Quantification dependents on predictions. In [12] complementary sensors are analyzed regarding advantages and weaknesses and compared with an early fusion approach. The results are shown over different artificial darkness levels and distance rings. Clear dependencies can be observed and are quantified as average precision values. In [11] decreasing performance of anchor-based approaches during dynamic situations like turning maneuvers, is reported. In the case of redundant systems, multiple predictions for potential objects can be available. Additional information can be obtained based on agreement or conflicts in the available predictions. In [2] the authors demonstrated that the combination of diverse detection approaches can lead to improved overall performance. Besides the dependencies induced by distance and darkness, indicators for the detection performance can be more diverse. Weather conditions like rain, snow, or fog can affect the system due to induced noise. Autonomous systems need to be aware of different situations and varying uncertainties. The expected uncertainties are depending on situational variations.

III. METHODS AND DATASET

In the following, the used detection systems and datasets are defined. Furthermore, the applied metrics for the performance evaluation are introduced.

A. Detection approaches and Dataset

This work utilizes detection systems implemented by [17]. Overall four detection systems based on lidar are involved. Used detection systems achitectures are pointpillar [10] and centerpoint [11] , using different pillar and voxel sizes. The test and training data is obtained from the nuScenes dataset [18]. The default split for the training and validation interval is discarded. A five fold cross validation scheme is applied as seen in Table I. Each detection approach is trained for all cross folds, resulting in five models. Each model predicts on eachs individual test fold. Therefore test results are obtained for the full dataset. For the validation data is divided into different situations. Weather situations are divided into four categories based on the provided scene description. Prediction cases are obtained afer matching predictions of multiple detection systems to a set of instances, compare Section III-B. This results in $(0,...2^n-1)$ cases where n is the number of detection systems and 0 represents not detected ground truth annotations. Predictions are divided into low and high distance level using median distance of available ground truth annotations as threshold.

B. Association of predictions

In order to compare the predictions of multiple approaches association is required. At a particular timestep multiple predictions are available representing multiple objects. Predictions of different detection systems are matched to a set of instances based on minimum center distance, as seen in Fig. 1. Similar to [16] geometric and semantic consistency are assumed. Therefore only predictions with same class and within geometric boundaries can be matched. Since the predictions are obtained after non-maximum uppression (NMS) only one prediction per detection system is assiciated with one instance. Predictions are assigned if a distance threshold of 2 m is not exceeded. Predictions with highest detecion score are matched first. If no association can be established a new instance is added to the existing population. An initial population of instances can be empty or based on prior information f. e. tracked object instances.

C. Applied metric

Results are obtained on instance level. A predicted instance is considered as true positive (tp) if at least one associated prediction is within the distance threshold. Results are provided for classes 'car' and 'pedestrians'. Average precision is calculated differentiating different situations as seen in Table II, Table III and Fig. 2. Precision over recall is calculated over accumulated range of predictions indicated by the detection score threshold. True-positive-rate over detection score is based on a window function with fixed size and no overlap. True-positive-rate can be interpreted as emperical probability $P(tp) = \frac{tp}{tp+fp} \in [0..1]$ given the detection score. Fusion results are obtained by evaluation of instances. The detection score of a particular instance is calculated by mean, min, or max values of the individual scores, without any prior knowledge.

IV. EXPERIMENTAL RESULTS

A. Distance dependency

To demonstrate distance dependencies, results are analyzed for different distance ranges as disscussed in Section III-A. It is expected that the performance decreases with increasing distance due to reduced point density. In Fig. 2 the precision recall curve as well as precison-detection-score relation for one detection system and one class is shown. A deecreased average precision at higher distance can be observed. Precision related to the detection score show still the same shape. Therefore it can be assumed that distance-based uncertainty is represented in the detection score. The detection approach is more confident at lower distance.

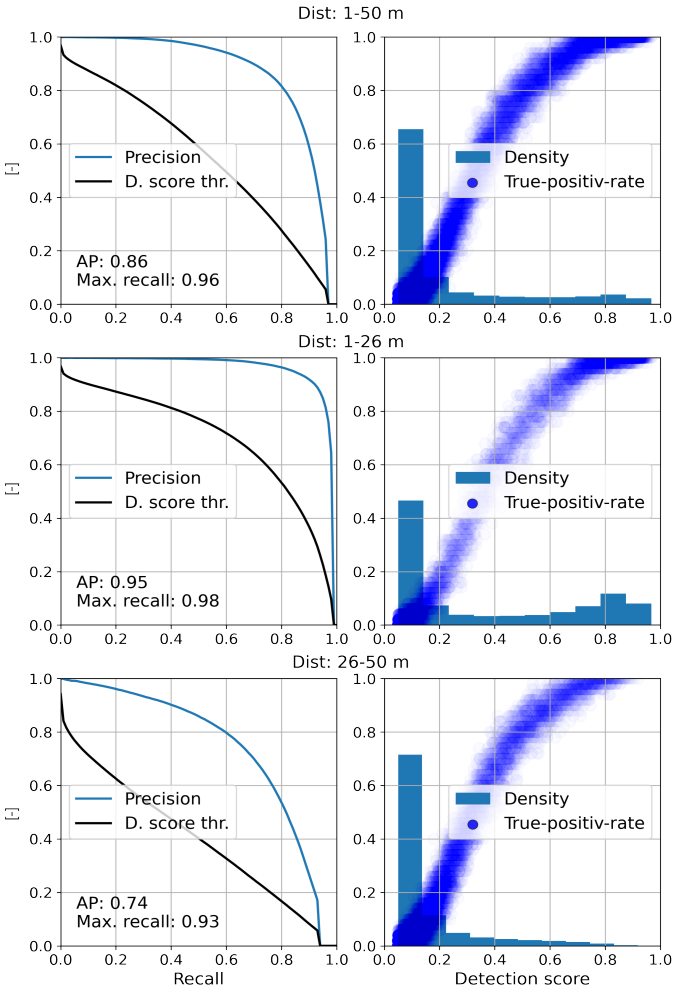

Fig. 2. Performance of PointPillar on class car at different distances. True-positive-rate is calculated with sample size of 50. Distance threshold is defined by median distance of available ground truth annotations.
Left: Precision-recall curve and associatet detection score threshold (D. score thr.).
Right: Precision over detection score and relativ densitiy over all given predictions.

B. Weather dependency

Performance dependencies are analyzed over four defined cases. Results are shown in Table II. Highest variation can

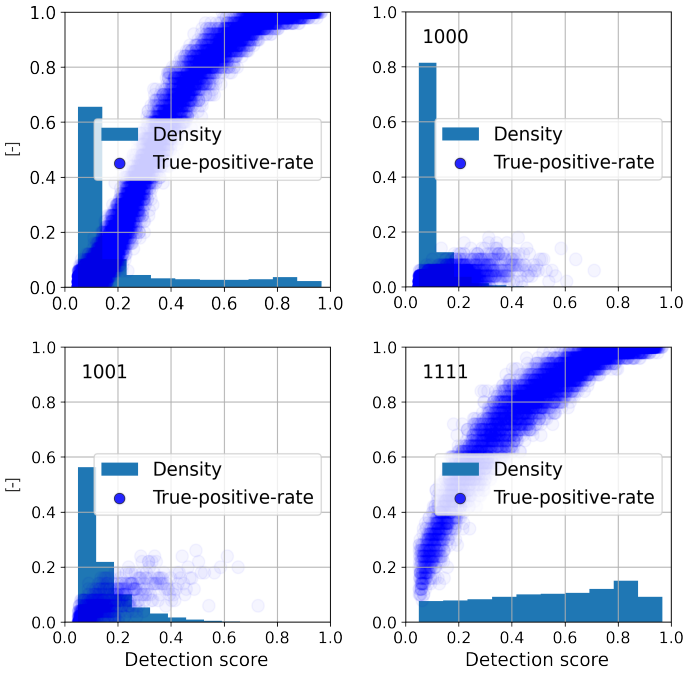

Fig. 3. Precision over detection score and relativ densitiy. Results shown for PointPillar on class car. True-positive-rate is calculated with sample size of 50.
Top left: All predictions of pointpillar.
Top right: Predictions detected by only PointPillar.
Bottom left: Predictions detected by PointPillar and CenterPoint Pillar 02.
Bottom right: Predictions detected by all detection systems.

be observed for rain and night (R & N). A possible reason is the small sample size for this case. The remaining cases show small variations but no significant dependency. It can be concluded that the used detection systems are robust against the testet weather effects. Never the less more specific cases like heavy rain, snow, etc. could be analyzed in future work.

C. Prediction cases and score fusion

Due to the initial matching predictions can be evaluated using information of other redundant systems. It can be assumed that instances with common detections are more likly to be correct. Furthermore the recall can be maximized since a single detection is sufficient to be considered in the evaluation. Predictions are divided into 2^n cases. Average precision is calculated for each case individually. Results are given for the particular detection system as well as for fused predictions. Results are shown in Table III. It can be observed that the achived performance increases if multiple detection systems detected an instance. Instances predicted by all four systems achive an average precision of 97.2 % while covering 85.7 % of all availabe annotations. The probability of a prediction beeing true positive changes significantly if other systems confirm the prediction. In Figure 3 it can be observed that depending on the detection case a different precision has to be expected. This can be utilized to isolate low probability predictions during inference. Evaluation of fused predictions shows improved AP of 2.9 % for class car and 4.36 % class pedestrian and demonstrates the potential of redundant detection systems.

V. CONCLUSION AND FUTURE WORK

In this contribution multiple lidar-based detecion approaches are analyzed with respect to situational dependencies and complementary performance potential. Results are obtained using cross fold validation and prediction matching. Variation in weather and light conditions show only insignificant variation in the performance and no clear dependency for the used dataset. Decrasing precision can be observed at higher distance level, however, this is already compensated by a lower confidence and is based on decreasing information density of the used data. The differentiation of the prediction cases enables the use of multiple detection systems and additional information. In this contribution, this is utilized by a fused detection score. Other fusion approaches and the use of additional information could lead to further improvements and should be be considered in future work.

ACKNOWLEDGMENT

We acknowledge support by the European Regional Development Fund (ERDF), grant-no. EFRE-0801714

REFERENCES

[1] *NHTSA Tesla Crash Preliminary Evaluation Report PE 16-007, U.S. Department of Transportation, National Highway Traffic Safety Administration, Technical report, Jan. 2017.*
[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.
[4] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.
[5] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.
[6] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3093–3097, 2019.
[7] C. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2017.
[8] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.
[9] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors (Basel, Switzerland)*, vol. 18, 2018.
[10] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 689–12 697, 2019.
[11] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3d object detection and tracking," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 779–11 788, 2021.
[12] E. R. Corral-Soto and B. Liu, "Understanding strengths and weaknesses of complementary sensor modalities in early fusion for object detection," *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1785–1792, 2020.
[13] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4603–4611, 2020.

TABLE II

AVERAGE PRECISION OF DIFFERENT DETECTION SYSTEMS UNDER DIFFERENT WEATHER SITUATIONS. BEST IN BOLD.

	Car					Pedestrian				
	Rain	Night	R & N	Default	All	Rain	Night	R & N	Default	All
PointPillar	85.05%	85.32%	88.59%	86.83%	86.45%	74.00%	72.22%	72.27%	75.22%	74.85%
CenterPoint Voxel 01	**86.10%**	**87.60%**	**90.74%**	**87.63%**	86.87%	**80.61%**	77.42%	72.62%	80.40%	80.01%
CenterPoint Voxel 0075	84.55%	86.86%	88.44%	86.21%	86.04%	80.53%	**80.45%**	**74.89%**	**82.13%**	81.56%
CenterPoint Pillar 02	81.26%	85.23%	88.61%	85.23%	84.23%	73.57%	72.03%	71.20%	73.20%	72.93%
Fused by min	-	-	-	-	89.45%	-	-	-	-	**84.37%**
Fused by mean	-	-	-	-	89.45%	-	-	-	-	**84.37%**
Fused by max	-	-	-	-	**89.76%**	-	-	-	-	83.07%

TABLE III

AVERAGE PRECISION OF DIFFERENT DETECTION SYSTEMS FOR DIFFERENT PREDICTION CASES. BEST IN BOLD.

	Car						Pedestrian					
			AP ?						AP ?			
case[1]	npos_rel[2]	mean[3]	#1000	#0100	#0010	#0001	npos_rel[2]	mean3	#1000	#0100	#0010	#0001
#0000	2.50%	-	-	-	-	-	1.80%	-	-	0.00%	-	-
#0001	0.10%	1.80%	-	-	-	1.80%	0.30%	0.70%	-	0.00%	-	0.70%
#0010	0.10%	6.70%	-	-	6.70%	-	0.30%	4.90%	-	0.00%	4.90%	-
#0011	0.10%	9.00%	-	-	**9.50%**	7.30%	0.20%	8.00%	-	0.00%	**8.70%**	4.00%
#0100	0.20%	5.80%	-	5.80%	-	-	0.40%	3.40%	-	3.40%	-	-
#0101	0.10%	14.30%	-	14.50%	-	9.90%	0.20%	5.80%	-	**7.30%**	-	3.50%
#0110	0.20%	**31.80%**	-	31.00%	24.00%	-	1.00%	**42.10%**	-	38.40%	33.60%	-
#0111	0.30%	**51.00%**	-	49.60%	47.70%	36.30%	1.60%	**56.10%**	-	52.70%	52.60%	25.40%
#1000	2.90%	3.30%	3.30%	-	-	-	1.80%	1.60%	1.60%	-	-	-
#1001	0.90%	**11.50%**	11.40%	-	-	8.20%	0.90%	5.90%	5.10%	-	-	3.10%
#1010	0.60%	**22.90%**	19.90%	-	15.90%	-	1.30%	**35.20%**	29.90%	-	25.10%	-
#1011	1.00%	31.60%	29.90%	-	27.10%	24.20%	1.70%	**47.10%**	42.20%	-	39.10%	27.80%
#1100	1.20%	**26.40%**	22.50%	20.30%	-	-	1.00%	16.30%	14.90%	10.50%	-	-
#1101	1.60%	**38.10%**	34.10%	34.90%	-	29.10%	1.60%	27.80%	28.10%	20.40%	-	16.50%
#1110	2.60%	**65.10%**	55.70%	58.30%	56.10%	-	4.50%	**62.20%**	52.50%	49.20%	56.40%	-
#1111	85.70%	**97.20%**	96.20%	96.80%	96.80%	96.20%	81.30%	**94.60%**	92.70%	92.50%	93.10%	90.40%

[1]Prediction case: ['PointPillar', 'CenterPoint Voxel 01', 'CenterPoint Voxel 0075', 'CenterPoint Pillar 02']

[2]Percentage of ground truth annotations covered by prediction case

[3]Instance is evaluated after fusing associated predictions by mean detection score

[14] D. Feng, Y. Cao, L. Rosenbaum, F. Timm, and K. C. J. Dietmayer, "Leveraging uncertainties for deep multi-modal object detection in autonomous driving," *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 877–884, 2020.

[15] C. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 918–927, 2018.

[16] S. Pang, D. D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10 386–10 393, 2020.

[17] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[18] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 618–11 628, 2020.